DOCUMENT RESUME

ED 237 539 TM 830 800

AUTHOR Prater, James M., Jr.
TITLE An Analysis of Selected Statistical Techniques Utilized in Quasi-Experimental Designs.
PUB DATE Nov 83
NOTE 34p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (12th, Nashville, TN, November 16-18, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Gains; *Analysis of Covariance; *Analysis of Variance; Computer Simulation; *Hypothesis Testing; Mathematical Models; Pretests Posttests; *Quasiexperimental Design; Reliability; Research Design; Research Methodology; Sample Size

ABSTRACT

The central purpose of the present study was to evaluate several statistical techniques (analysis of variance (ANOVA) on raw gains, standardized ANOVA, standard analysis of covariance (ANCOVA), and z-score ANOVA) under various quasi-experimental conditions (three levels of reliability, three levels of sample size, three levels of gain, and the equality of pretest means). One thousand sets of data were generated for each combination of statistical technique/conditions. Based on the results of the study, none of the statistical techniques were found to appropriately model reality. There was no statistically significant difference between the results produced by all four statistical techniques when there was no gain in either group and the pretest means of both groups were equal. ANOVA on raw gains and z-score ANOVA produced similar results when there was no gain in either group and the pretest means were different for both groups. The idea of developing a multiple analysis scheme was not supported for any of the statistical techniques used in this study. Finally, the study did not support the use of the statistical techniques investigated under the quasi-conditions constructed for this study. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED237539

An Analysis of Selected Statistical Techniques Utilized in Quasi-Experimental Designs

James M. Prater, Jr.

The University of Alabama

1801 17th Avenue, East
Tuscaloosa, Alabama 35404
Phone: 205/556-4673

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. M. Prater, Jr.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the

Annual Meeting

of the

Mid-South Educational Research Association

November 16-18, 1983

TM 830 802

AN ANALYSIS OF SELECTED STATISTICAL TECHNIQUES UTILIZED IN QUASI-EXPERIMENTAL DESIGNS

Introduction

Almost all statistical theoreticians agree that random assignment of subjects to groups is the most essential element of a good research design. However, in most educational settings, a truly randomized experiment is difficult to undertake, and many researchers opt for a quasi-experimental design. Generally, the quasi-experiment entails nonrandom selection and assignment of individuals or groups to treatment conditions being investigated.

Statement of the Problem

Nonrandom assignment of subjects in quasi-experimental designs presents many problems for researchers who are attempting to evaluate the effect that a certain treatment has (as opposed to a more traditional learning program) in improving the student's ability to learn (Bennett & Lumsdaine, 1975; Harris, 1963; Huck & McLean, 1975; McKinlay, 1975; Neel, 1970).

The central purpose of the present study was to evaluate several statistical techniques (i.e., ANOVA on raw gains, standardized ANOVA, standard ANCOVA, and z-score ANOVA) under various quasi-experimental conditions (i.e., three levels of reliability, three levels of sample size, three levels of gain, and the equality of pretest means). The conditions used in this study were defined in terms of the factors of reliability, sample size, gain, and equality of pretest means. Subsequently, 1,000 sets of data were generated for

each combination of statistical technique-conditions which resulted in generating 216,000 sets of data combinations of statistical technique-factor sets investigated.

Null Hypotheses

1. There will be no difference between the expected and observed error rates for each selected statistical technique (i.e. ANOVA on raw gains, standardized ANOVA, standard ANCOVA, and z-score ANOVA) under each combination of factors (i.e., level of reliability, sample size, gain, and equality of pretest means).

2. For each condition of the factors (i.e., levels of reliability, sample size, gain, and equality of pretest means) where the error rates from two or more statistical techniques are not significantly different than expected, there is no difference among the selected statistical techniques.

3. The levels of reliability, sample size, gain, and equality of pretest means are not separately nor in combination related to the observed error rates for each statistical technique (i.e., ANOVA on raw gains, standardized ANOVA, standard ANCOVA, and z-score ANOVA).

Significance of the Problem

This study was designed to make use of simulation methods in order to evaluate suggested techniques purported to reduce errors associated with quasi-experimental designs. While many authors have supported one technique over another, the general consensus among many researchers (Cook & Campbell, 1976; Cook & Reichardt, 1976; Cronbach, Rogosa, Price, & Floden, 1976; Porter & Chibucos, 1974, 1975) is formulation of a multiple analysis strategy when using quasi-experimental designs. By using multiple analyses, one should be able to rule out different threats to internal validity with different analyses

and to establish confidence intervals for estimates of important parameters that are not known.

Definition of Terms

Experimental design is defined as

> one or more treatments (programs) that are administered to some set of persons (or other units) drawn at random from a specified population; and that observations (or measurements) are made to learn how (or how much) some relevant aspect of their behavior following treatment differs from like behavior on the part of an untreated or control group also drawn at random from the same population. (Riecken & Boruch, 1974, p. 3)

Quasi-experimental designs are defined as

> natural social settings in which the research person can introduce something like experimental design into his scheduling of data collection procedures, even though he lacks the full control over the scheduling of experimental stimuli, which makes a true experiment possible. (Campbell & Stanley, 1963, p. 34)

Random numbers are defined as

> numbers generated by a completely deterministic arithmetical process; the resulting set of numbers has various statistical properties that together are called randomness. (Scheid, 1968, p. 401)

Monte Carlo simulation refers to

> methods of providing arithmetical imitations of real phenomena. In the broad sense, this describes the general idea of applied mathematics. A differential equation may, for example, simulate the flight of a missile. Here, however, the term simulation refers to the imitation of random processes by Monte Carlo methods. The classic example is the simulation of a neutron's motion into a reactor wall, its zigzag path being imitated by an arithmetical random walk. (Scheid, 1968, p. 401)

Temporal erosion is the

> shrinkage of test-retest correlations as the time lapses between the administration of the test and retest increases. (Campbell & Boruch, 1975)

Fan spread pattern occurs

> when the lower-scoring pretest group falls further behind the higher-scoring group over time. (Kenny, 1975)

Internal validity concerns

> the efficiency with which an experiment controls extraneous variables that include history, maturation, testing instrumentation, statistical regression, differential selection, experimental mortality, and selection-maturation interaction. (Campbell & Stanley, 1963)

Attenuation occurs when:

> the pretest and posttest are correlated resulting in a lower coefficient that is due to the unreliability of one or both tests. (Borg & Gall, 1971)

Assumptions

1. The numerical generation technique will produce random numbers.

2. The empirical population distribution generated will be normally distributed.

3. The reliability between the pretest and posttest will reflect actual reliability coefficients found in educational settings.

Limitations

This study was designed to evaluate the effectiveness of four statistical techniques in analyzing a quasi-experimental design. Other statistical techniques were not included in this study because critiques or review of these other techniques in the literature indicated they were not suitable for this type of design. The combinations of differing reliabilities, sample sizes, mean gains, and pretest means have been preset to allow for reliable assessments of their relative effectiveness and yield the best combinations to be used in future research involving this type of design. The study was also limited to two groups with reliabilities and gain scores arranged in comparable increments, so as to ensure the most effective evaluation possible.

Review of Related Literature

The most common design used in educational and evaluation research involves a comparison of a nonrandom group of students to another nonrandom group receiving a particular treatment (or no treatment) where the group's performance is measured both before and after the treatment has been administered. In a 1963 publication, Campbell and Stanley referred to this procedure as the nonequivalent control group design. Cook and Campbell in 1976 referred to this same method as a form of quasi-experimental design. This particular design has two inherent problems: (a) control of threats to internal validity and (b) control of measurement error. Despite these inherent problems, wide use of quasi-experimental designs in educational and evaluation research, as opposed to true experiments, results from the high cost of implementing a random experiment, reluctance of local school administration to aid in implementation of a laborious process, and the ethical problems when a promising new innovation is restricted to a randomly selected group (Coulson, 1978). These very practical problems of quasi-experiments make it essential that we turn our attention to the strengths and weaknesses of those statistical techniques now being utilized in quasi-experiments.

Lord (1960) warned that simple analysis of covariance could not remove all the bias introduced by using a quasi-experimental design. He stated that analysis of covariance fails to adjust for the attenuation of regression coefficients (due to pretest measurement error) on which estimates of the treatment's effect are based. Lord provided a method for adjusting pretest unreliability when assumptions about the structure of the test and nature of the measurement error can be accepted. In a second paper, Lord (1967) compared simple gain score analysis and analysis of covariance. He found that the two techniques provided different results when applied to quasi-experimental designs.

Lord concluded that neither of these two statistical techniques conforms to reality and no statistical procedure can logically adjust for preexisting differences between groups (Lord, 1969).

Campbell and Stanley (1963) noted that two threats to internal validity are selection-maturation and statistical regression. They imply that analysis of covariance and matching cannot control these two threats satisfactorily, and simple gains score analysis makes no allowances for either threat. Porter (1967) developed an adjustment procedure for analysis of covariance which intended to solve the problem of an unreliable pretest. However, McLean (1974) and McLean, Ware, and McClave (1975) tested both simple analysis of covariance and Porter's adjusted analysis of covariance and concluded that both techniques produce erroneous results.

Porter and Chibucos (1974, 1975) in two papers compared analysis of variance, gain score analysis, analysis of covariance, and reliability-adjusted analysis of covariance in both random experiments and quasi-experiments. The results suggested that the reliability-adjusted analysis of covariance is the more effective method to be used in quasi-experiments when the assumption of measurement error provides the only reason for the pretest's being an unreliable covariate.

Kenney (1975) dealt with three methods of selecting treatment groups for quasi-experimental designs. The third method was the most common case where assignment was based on existing social groups that were assumed to be constantly changing in a fan-spread pattern. For the third method, he suggested that a standardized change score analysis be utilized.

Hendrix, Carter, and Hintze (1978) compared five methods of analyzing quasi-experimental designs which included: an analysis of variance on the posttest scores, an analysis of covariance on the raw gain scores, an analysis

of variance with a pretest-posttest factor included in the model, an analysis of covariance using the pretest as a covariate, and an analysis of covariance on the gain scores using the pretest as a covariate. The authors found that the analysis of covariance on gain scores using the pretest as the covariate was the most desirable procedure.

The z-score method is used primarily in epidemiologic research. This method used a covariate(s) in order to determine which grouping scheme should be employed. Once the grouping scheme was decided, the treatment and control groups were subdivided into smaller groups made up of elements from the two larger groups. Means and standard deviations were then used to standardize each of the subgroups using the standard normal variable z. After standardization was completed, the subgroup elements were returned to their original groups and statistical analysis was then used in order to determine treatment-effect. The major drawback of the z-score method was the somewhat arbitrary method for selecting the subgroups (Kleinbaum & Kupper, 1978).

Procedures

Since quasi-experimental designs are most often employed in educational and evaluation research, investigation of threats to internal validity and measurement error leading to development of useful decision strategies should be continuously explored. The method of Monte Carlo simulation was used to select the best combination of four statistical techniques now being used, or proposed for use, when a quasi-experimental design is encountered. The four statistical methods evaluated were ANOVA on raw gains, standardized ANOVA, standard ANCOVA, and z-score ANOVA.

Design of the Study

The study evaluated four statistical techniques in combination with four

factors which comprised 54 sets of conditions. These factors included reliability (levels: .50, .70, .90), sample size (levels: 10, 55, 100), gain (levels: no gain in either group, gain in favor of the treatment group, and gain in favor of the no-treatment group), and the equality of pretest means. For convenience, the design of the study is illustrated in Table 1. The three levels of reliability were selected to detect problems in the selected statistical techniques when conditions of poor reliability, moderate reliability, and high reliability are present (Helmstadter, 1964). The corresponding sample sizes were selected to detect problems in the selected statistical techniques when small samples, moderate samples, and large samples are used (Ott, Mendenhall, & Larson, 1978). The levels of gain were selected in order to reflect a realistic educational setting (McLean, 1974). The mean values were also selected to reflect a realistic educational setting (McLean, 1974). Thus, the true score means for both groups were preset at 100 when both groups have equal means. When the treatment group had a lower mean, the mean was preset at 80 (McLean, 1974).

Data Generation Technique

The data generation technique used in this study was obtained from the MATH-PACK computer program package (UNIVAC, 1970) designed for use on the UNIVAC computer. The program RANDN (UNIVAC, 1970) always produces a normal distribution with a specified mean and standard deviation.

The input value submitted to RANDN called for generation of a normal distribution with mean zero and unit variance. The variables were then transformed to have the desired means and variances for the selected analyses performed (McLean, 1974). The normality of the data generated was tested using the Kolmogorov-Smirnov goodness of fit test (Gibbons, 1976).

TABLE 1

DESIGN OF THE STUDY

Situations					Situations				
No.	PT$\bar{X}$'s	r	n	Gain	No.	PT$\bar{X}$'s	r	n	Gain
1	$\bar{X}_1 = \bar{X}_2$	.50	10	GT	28	$\bar{X}_1 = \bar{X}_2$	.50	10	NG
2	$\bar{X}_1 = \bar{X}_2$	.50	55	GT	29	$\bar{X}_1 = \bar{X}_2$	.50	55	NG
3	$\bar{X}_1 = \bar{X}_2$	.50	100	GT	30	$\bar{X}_1 = \bar{X}_2$	.50	100	NG
4	$\bar{X}_1 = \bar{X}_2$	.70	10	GT	31	$\bar{X}_1 = \bar{X}_2$	.70	10	NG
5	$\bar{X}_1 = \bar{X}_2$	.70	55	GT	32	$\bar{X}_1 = \bar{X}_2$	.70	55	NG
6	$\bar{X}_1 = \bar{X}_2$	.70	100	GT	33	$\bar{X}_1 = \bar{X}_2$	.70	100	NG
7	$\bar{X}_1 = \bar{X}_2$	.90	10	GT	34	$\bar{X}_1 = \bar{X}_2$	.90	10	NG
8	$\bar{X}_1 = \bar{X}_2$	.90	55	GT	35	$\bar{X}_1 = \bar{X}_2$	.90	55	NG
9	$\bar{X}_1 = \bar{X}_2$	.90	100	GT	36	$\bar{X}_1 = \bar{X}_2$	.90	100	NG
10	$\bar{X}_1 > \bar{X}_2$	.50	10	GT	37	$\bar{X}_1 > \bar{X}_2$	.50	10	NG
11	$\bar{X}_1 > \bar{X}_2$	.50	55	GT	38	$\bar{X}_1 > \bar{X}_2$	.50	55	NG
12	$\bar{X}_1 > \bar{X}_2$	.50	100	GT	39	$\bar{X}_1 > \bar{X}_2$	.50	100	NG
13	$\bar{X}_1 > \bar{X}_2$	.70	10	GT	40	$\bar{X}_1 > \bar{X}_2$	.70	10	NG
14	$\bar{X}_1 > \bar{X}_2$	.70	55	GT	41	$\bar{X}_1 > \bar{X}_2$	.70	55	NG
15	$\bar{X}_1 > \bar{X}_2$	.70	100	GT	42	$\bar{X}_1 > \bar{X}_2$	.70	100	NG
16	$\bar{X}_1 > \bar{X}_2$	.90	10	GT	43	$\bar{X}_1 > \bar{X}_2$	.90	10	NG
17	$\bar{X}_1 > \bar{X}_2$	.90	55	GT	44	$\bar{X}_1 > \bar{X}_2$	.90	55	NG
18	$\bar{X}_1 > \bar{X}_2$	.90	100	GT	45	$\bar{X}_1 > \bar{X}_2$	.90	100	NG
19	$\bar{X}_2 > \bar{X}_1$	.50	10	GNT	46	$\bar{X}_2 > \bar{X}_1$	.50	10	NG
20	$\bar{X}_2 > \bar{X}_1$	.50	55	GNT	47	$\bar{X}_2 > \bar{X}_1$	.50	55	NG
21	$\bar{X}_2 > \bar{X}_1$	.50	100	GNT	48	$\bar{X}_2 > \bar{X}_1$	.50	100	NG
22	$\bar{X}_2 > \bar{X}_1$	.70	10	GNT	49	$\bar{X}_2 > \bar{X}_1$	.70	10	NG
23	$\bar{X}_2 > \bar{X}_1$	.70	55	GNT	50	$\bar{X}_2 > \bar{X}_1$	.70	55	NG
24	$\bar{X}_2 > \bar{X}_1$	.70	100	GNT	51	$\bar{X}_2 > \bar{X}_1$	.70	100	NG
25	$\bar{X}_2 > \bar{X}_1$	.90	10	GNT	52	$\bar{X}_2 > \bar{X}_1$	.90	10	NG
26	$\bar{X}_2 > \bar{X}_1$	.90	55	GNT	53	$\bar{X}_2 > \bar{X}_1$	.90	55	NG
27	$\bar{X}_2 > \bar{X}_1$	.90	100	GNT	54	$\bar{X}_2 > \bar{X}_1$	.90	100	NG

NG = no gain

PT$\bar{X}$'s = pretest means

$\bar{X}_1$ = treatment group mean

$\bar{X}_2$ = no-treatment group

GT = gain in treatment group

GNT = gain in no-treatment group

r = reliability

n = sample size

Testing the Null Hypotheses

The proportion of rejections for each statistical technique-factor set was the basis for testing the first null hypothesis. The proportion of rejections of the 1,000 samples provided empirical estimates of Type I and Type II errors for each statistical technique-factor set. Because of the large number of individual tests of hypothesis (216), the experiment-wise error-rate was set at .10. Thus, the hypothesis-wise error-rate was determined to be .0005. Each of the 54 statistical technique-factor sets was then evaluated using the z statistic.

The second null hypothesis compared those statistical techniques found to be effective when tested as part of the first null hypothesis for each of the 54 statistical technique-factor sets. Once those techniques were determined, the subsequent statistical technique-factor sets where two or more statistical techniques were found to be effective were compared further. The focus centered around comparing the proportions of rejections among the effective statistical techniques for each statistical technique-factor set using the chi square statistic.

The third null hypothesis tested the levels of reliability, sample size, and equality of pretest means in a series of eight factorial designs where the generated error rates (i.e., alphas and betas) were used as the criterion variables for each of the statistical techniques. Significant interactions between the tested factors provided evidence supporting the rejection of the null hypothesis.

Data Analysis

This section contains the results obtained in the study. The results were based on the previously determined procedures. The results of testing each null

hypothesis are presented in order beginning with Null Hypothesis 1.

The first null hypothesis dealt with the effectiveness of each statistical technique (standard ANCOVA, ANOVA on raw gains, standardized ANOVA, and z-score ANOVA) to control Type I and Type II errors when combined with selected factors (i.e., level of reliability, sample size, gain, and equality of pretest means). A z-test was used to compare the observed fraction of significant F's to preset alpha and beta values. The results of the z-tests are recorded in Tables 2 through 9.

Table 2 presents the observed fraction of significant F's when there was a gain in either group. Inspection of the table reveals that out of the 27 conditions tested, only two conditions were found to be nonsignificant. This amounted to rejecting 25 null hypotheses resulting in a 99.95% confidence that the subsequent alternative hypotheses were correct. Table 3 presents the observed fraction of significant F's when there was no gain in either group. Inspection of the table reveals that out of 27 conditions tested only nine conditions were found to be nonsignificant. The findings indicate standard analysis of covariance did not provide appropriate control of Type II errors. However, the statistical technique did provide appropriate control of Type I errors when both pretest means were equal regardless of the reliability of the pretest of sample size used.

Table 4 presents the observed fraction of significant F's using analysis of variance on raw gains when there was a gain in either group. Inspection of the table reveals all 27 conditions were found to be statistically significant at the .0005 level. This amounted to rejecting all 27 null hypotheses resulting in a 99.95% confidence that the subsequent alternative hypotheses were correct. Table 5 presents the observed fraction of significant F's using analysis of variance on raw gains when there was no gain in either group. The findings

TABLE 2
FRACTION OF SIGNIFICANT F's USING STANDARD ANALYSIS
OF COVARIANCE AND THERE WAS A GAIN IN EITHER GROUP

Reliability of Treatment Group Scores	Reliability of No-Treatment Group Scores	Sample Size of Each Group	Size of Gain	Pretest Mean of Treatment Group	Pretest Mean of No-Treatment Group	Fraction of Significant [a] F's	[b] Computed Z-Values
.50	.50	10	1.88	100	100	.074	-26.962
.50	.50	55	.75	100	100	.074	-26.962
.50	.50	100	.56	100	100	.069	-27.278
.70	.70	10	1.88	100	100	.090	-25.949
.70	.70	55	.75	100	100	.054	-28.228
.70	.70	100	.56	100	100	.091	-25.886
.90	.90	10	1.88	100	100	.123	-23.861
.90	.90	55	.75	100	100	.111	-24.620
.90	.90	100	.56	100	100	.130	-23.418
.50	.50	10	1.88	100	80	.657	9.937
.50	.50	55	.75	100	80	.994	31.266
.50	.50	100	.56	100	80	1.000	31.646
.70	.70	10	1.88	100	80	.485	-.949
.70	.70	55	.75	100	80	.927	27.025
.70	.70	100	.56	100	80	.999	31.582
.90	.90	10	1.88	100	80	.316	-11.646
.90	.90	55	.75	100	80	.553	3.354
.90	.90	100	.56	100	80	.794	18.608
.50	.50	10	1.88	80	100	.414	-5.443
.50	.50	55	.75	80	100	.976	30.128
.50	.50	100	.56	80	100	1.000	31.646
.70	.70	10	1.88	80	100	.239	-16.619
.70	.70	55	.75	80	100	.795	18.709
.70	.70	100	.56	80	100	.964	30.633
.90	.90	10	1.88	80	100	.074	-26.962
.90	.90	55	.75	80	100	.159	-21.582
.90	.90	100	.56	80	100	.374	-7.975

[a] Monte Carlo derived beta values.

[b] Z test is significant if computed Z value is equal to the absolute value of 3.475.

TABLE 3
FRACTION OF SIGNIFICANT F's USING STANDARD ANALYSIS
OF COVARIANCE AND THERE WAS NO GAIN IN EITHER GROUP

Reliability of Treatment Group Scores	Reliability of No-Treatment Group Scores	Sample Size of Each Group	Size of Gain	Pretest Mean of Treatment Group	Pretest Mean of No-Treatment Group	Fraction of Significant [a] F's	[b] Computed Z-Values
.50	.50	10	0.0	100	100	.052	.290
.50	.50	55	0.0	100	100	.031	-2.750
.50	.50	100	0.0	100	100	.055	.725
.70	.70	10	0.0	100	100	.057	1.014
.70	.70	55	0.0	100	100	.047	-.435
.70	.70	100	0.0	100	100	.058	1.159
.90	.90	10	0.0	100	100	.046	-.580
.90	.90	55	0.0	100	100	.047	-.435
.90	.90	100	0.0	100	100	.045	-.725
.50	.50	10	0.0	100	80	.517	67.681
.50	.50	55	0.0	100	80	.992	136.523
.50	.50	100	0.0	100	80	1.000	137.681
.70	.70	10	0.0	100	80	.330	40.580
.70	.70	55	0.0	100	80	.882	120.580
.70	.70	100	0.0	100	80	.995	136.956
.90	.90	10	0.0	100	80	.143	13.478
.90	.90	55	0.0	100	80	.335	41.304
.90	.90	100	0.0	100	80	.595	78.986
.50	.50	10	0.0	80	100	.511	66.812
.50	.50	55	0.0	80	100	.998	137.391
.50	.50	100	0.0	80	100	1.000	137.681
.70	.70	10	0.0	80	100	.346	42.898
.70	.70	55	0.0	80	100	.857	116.956
.70	.70	100	0.0	80	100	.986	135.652
.90	.90	10	0.0	80	100	.148	14.029
.90	.90	55	0.0	80	100	.315	38.406
.90	.90	100	0.0	80	100	.578	76.522

[a] Monte Carlo derived alpha value.

[b] Z test is significant if computed Z value is equal to the absolute value of 3.475.

TABLE 4

FRACTION OF SIGNIFICANT F's USING ANALYSIS OF VARIANCE ON RAW GAINS AND THERE WAS GAIN IN EITHER GROUP

Reliability of Treatment Group Scores	Reliability of No-Treatment Group Scores	Sample Size of Each Group	Size of Gain	Pretest Mean of Treatment Group	Pretest Mean of No-Treatment Group	Fraction of Significant F's[a]	Computed Z-Values[b]
.50	.50	10	1.88	100	100	.054	-28.228
.50	.50	55	.75	100	100	.070	-27.152
.50	.50	100	.56	100	100	.068	-27.342
.70	.70	10	1.88	100	100	.089	-26.013
.70	.70	55	.75	100	100	.077	-26.772
.70	.70	100	.56	100	100	.095	-25.633
.90	.90	10	1.88	100	100	.118	-24.177
.90	.90	55	.75	100	100	.114	-24.430
.90	.90	100	.56	100	100	.123	-23.861
.50	.50	10	1.88	100	80	.063	-27.658
.50	.50	55	.75	100	80	.059	-27.911
.50	.50	100	.56	100	80	.073	-27.025
.70	.70	10	1.88	100	80	.081	-26.519
.70	.70	55	.75	100	80	.077	-26.772
.70	.70	100	.56	100	80	.083	-26.392
.90	.90	10	1.88	100	80	.137	-22.975
.90	.90	55	.75	100	80	.121	-23.987
.90	.90	100	.56	100	80	.121	-23.987
.50	.50	10	1.88	80	100	.072	-27.089
.50	.50	55	.75	80	100	.052	-28.354
.50	.50	100	.56	80	100	.063	-27.658
.70	.70	10	1.88	80	100	.069	-27.278
.70	.70	55	.75	80	100	.064	-27.595
.70	.70	100	.56	80	100	.090	-25.949
.90	.90	10	1.88	80	100	.120	-24.051
.90	.90	55	.75	80	100	.120	-24.051
.90	.90	100	.56	80	100	.129	-23.481

[a] Monte Carlo derived beta values.

[b] Z test is significant if computed Z value is equal to the absolute value of 3.475.

TABLE 5
FRACTION OF SIGNIFICANT F's USING ANALYSIS OF
VARIANCE ON RAW GAINS AND THERE WAS NO GAIN IN EITHER GROUP

Reliability of Treatment Group Scores	Reliability of No-Treatment Group Scores	Sample Size of Each Group	Size of Gain	Pretest Mean of Treatment Group	Pretest Mean of No-Treatment Group	Fraction of Significant F's[a]	Computed Z-Values[b]
.50	.50	10	0.0	100	100	.052	.290
.50	.50	55	0.0	100	100	.050	.000
.50	.50	100	0.0	100	100	.061	1.594
.70	.70	10	0.0	100	100	.054	.598
.70	.70	55	0.0	100	100	.052	.290
.70	.70	100	0.0	100	100	.041	-1.304
.90	.90	10	0.0	100	100	.053	.435
.90	.90	55	0.0	100	100	.050	.000
.90	.90	100	0.0	100	100	.048	-.290
.50	.50	10	0.0	100	80	.049	-.145
.50	.50	55	0.0	100	80	.040	-1.449
.50	.50	100	0.0	100	80	.051	.145
.70	.70	10	0.0	100	80	.042	-1.159
.70	.70	55	0.0	100	80	.057	1.014
.70	.70	100	0.0	100	80	.050	.000
.90	.90	10	0.0	100	80	.043	-1.014
.90	.90	55	0.0	100	80	.046	-.580
.90	.90	100	0.0	100	80	.059	1.304
.50	.50	10	0.0	80	100	.051	.145
.50	.50	55	0.0	80	100	.057	1.014
.50	.50	100	0.0	80	100	.055	.725
.70	.70	10	0.0	80	100	.035	-2.174
.70	.70	55	0.0	80	100	.050	.000
.70	.70	100	0.0	80	100	.050	.000
.90	.90	10	0.0	80	100	.040	-1.449
.90	.90	55	0.0	80	100	.059	1.304
.90	.90	100	0.0	80	100	.051	.145

[a] Monte Carlo derived alpha value.

[b] Z test is significant if computed Z value is equal to the absolute value of 3.475.

indicate analysis of variance on raw gains did not provide appropriate control of Type II errors. However, the statistical technique did provide appropriate control of Type I errors in all 27 conditions where there was no gain in either group.

Table 6 presents the observed fraction of significant F's using standardized analysis of variance when there was no gain in either group. Inspection of the table reveals all 27 conditions were found to be statistically significant at the .0005 level. Table 7 presents the observed fraction of significant F's using standardized analysis of variance when there was no gain in either group. Inspection of the table reveals that out of 27 conditions tested only nine conditions were found to be nonsignificant. Thus, enough statistical evidence was presented to indicate standardized analysis of variance did not provide appropriate control of Type II errors. However, the statistical technique did provide appropriate control of Type I errors when both pretest means were equal regardless of the reliability of the pretest or sample size used.

Table 8 presents the observed fractions of significant F's using z-score analysis of variance when there was no gain in either group. Inspection of the table reveals all 27 conditions were found to be statistically significant at the .0005 level. Table 9 presents the observed fraction of significant F's using z-score analysis of variance when there was no gain in either group. The results indicated z-score analysis of variance did not provide appropriate control of Type II errors. However, the statistical technique did provide appropriate control of Type I errors in all but three conditions where there was no gain in either group.

The second null hypothesis was based on the results obtained from the testing of the first null hypothesis. Once the statistical technique-factor combination sets were determined, they were compared using the chi-square test in order to disclose if there was a statistically significant difference in the observed

TABLE 6
FRACTION OF SIGNIFICANT F's USING STANDARDIZED
ANALYSIS OF VARIANCE AND THERE WAS A GAIN IN EITHER GROUP

Reliability of Treatment Group Scores	Reliability of No-Treatment Group Scores	Sample Size of Each Group	Size of Gain	Pretest Mean of Treatment Group	Pretest Mean of No-Treatment Group	Fraction of Significant F's[a]	Computed Z-Values[b]
.50	.50	10	1.88	100	100	.073	-27.025
.50	.50	55	.75	100	100	.072	-27.089
.50	.50	100	.56	100	100	.072	-27.089
.70	.70	10	1.88	100	100	.078	-26.709
.70	.70	55	.75	100	100	.061	-27.785
.70	.70	100	.56	100	100	.078	-26.709
.90	.90	10	1.88	100	100	.116	-24.304
.90	.90	55	.75	100	100	.108	-24.810
.90	.90	100	.56	100	100	.110	-24.684
.50	.50	10	1.88	100	80	.004	-31.392
.50	.50	55	.75	100	80	.000	-31.456
.50	.50	100	.56	100	80	.003	-31.456
.70	.70	10	1.88	100	80	.001	-31.582
.70	.70	55	.75	100	80	.002	-31.519
.70	.70	100	.56	100	80	.001	-31.582
.90	.90	10	1.88	100	80	.030	-29.747
.90	.90	55	.75	100	80	.001	-31.582
.90	.90	100	.56	100	80	.001	-31.582
.50	.50	10	1.88	80	100	.002	-31.519
.50	.50	55	.75	80	100	.004	-31.392
.50	.50	100	.56	80	100	.006	-31.266
.70	.70	10	1.88	80	100	.006	-31.266
.70	.70	55	.75	80	100	.006	-31.266
.70	.70	100	.56	80	100	.006	-31.266
.90	.90	10	1.88	80	100	.012	-30.886
.90	.90	55	.75	80	100	.029	-29.810
.90	.90	100	.56	80	100	.029	-29.810

[a] Monte Carlo derived beta values.

[b] Z test is significant if computed Z value is equal to the absolute value of 3.475.

TABLE 7
FRACTION OF SIGNIFICANT F's USING STANDARDIZED
ANALYSIS OF VARIANCE AND THERE WAS NO GAIN IN EITHER GROUP

Reliability of Treatment Group Scores	Reliability of No-Treatment Group Scores	Sample Size of Each Group	Size of Gain	Pretest Mean of Treatment Group	Pretest Mean of No-Treatment Group	Fraction of Significant [a] F's	[b] Computed Z-Values
.50	.50	10	0.0	100	100	.048	-.290
.50	.50	55	0.0	100	100	.052	.290
.50	.50	100	0.0	100	100	.057	1.014
.70	.70	10	0.0	100	100	.041	-1.304
.70	.70	55	0.0	100	100	.055	.725
.70	.70	100	0.0	100	100	.052	.290
.90	.90	10	0.0	100	100	.064	2.029
.90	.90	55	0.0	100	100	.031	-2.754
.90	.90	100	0.0	100	100	.052	.290
.50	.50	10	0.0	100	80	.000	-7.246
.50	.50	55	0.0	100	80	.001	-7.102
.50	.50	100	0.0	100	80	.001	-7.102
.70	.70	10	0.0	100	80	.001	-7.102
.70	.70	55	0.0	100	80	.000	-7.246
.70	.70	100	0.0	100	80	.003	-6.812
.90	.90	10	0.0	100	80	.001	-7.102
.90	.90	55	0.0	100	80	.005	-6.522
.90	.90	100	0.0	100	80	.006	-6.377
.50	.50	10	0.0	80	100	.000	-7.246
.50	.50	55	0.0	80	100	.002	-6.956
.50	.50	100	0.0	80	100	.002	-6.956
.70	.70	10	0.0	80	100	.000	-7.246
.70	.70	55	0.0	80	100	.003	-6.812
.70	.70	100	0.0	80	100	.004	-6.667
.90	.90	10	0.0	80	100	.001	-7.102
.90	.90	55	0.0	80	100	.006	-6.377
.90	.90	100	0.0	80	100	.006	-6.377

[a] Monte Carlo derived alpha value.

[b] Z test is significant if computed Z value is equal to the absolute value of 3.475.

TABLE 8

FRACTION OF SIGNIFICANT F's USING Z-SCORE ANALYSIS OF VARIANCE AND THERE WAS A GAIN IN EITHER GROUP

Reliability of Treatment Group Scores	Reliability of No-Treatment Group Scores	Sample Size of Each Group	Size of Gain	Pretest Mean of Treatment Group	Pretest Mean of No-Treatment Group	Fraction of Significant [a]F's	[b]Computed Z-Values
.50	.50	10	1.88	100	100	.072	-27.089
.50	.50	55	.75	100	100	.070	-27.152
.50	.50	100	.56	100	100	.071	-27.152
.70	.70	10	1.88	100	100	.075	-26.899
.70	.70	55	.75	100	100	.083	-26.392
.70	.70	100	.56	100	100	.089	-26.013
.90	.90	10	1.88	100	100	.128	-23.544
.90	.90	55	.75	100	100	.127	-23.608
.90	.90	100	.56	100	100	.131	-23.354
.50	.50	10	1.88	100	80	.060	-27.848
.50	.50	55	.75	100	80	.058	-27.975
.50	.50	100	.56	100	80	.081	-26.519
.70	.70	10	1.88	100	80	.099	-25.380
.70	.70	55	.75	100	80	.081	-26.519
.70	.70	100	.56	100	80	.090	-25.949
.90	.90	10	1.88	100	80	.132	-23.291
.90	.90	55	.75	100	80	.137	-22.975
.90	.90	100	.56	100	80	.117	-24.240
.50	.50	10	1.88	80	100	.069	-27.278
.50	.50	55	.75	80	100	.070	-27.215
.50	.50	100	.56	80	100	.076	-26.835
.70	.70	10	1.88	80	100	.082	-26.456
.70	.70	55	.75	80	100	.069	-27.278
.70	.70	100	.56	80	100	.081	-26.519
.90	.90	10	1.88	80	100	.132	-23.291
.90	.90	55	.75	80	100	.137	-22.975
.90	.90	100	.56	80	100	.115	-24.367

[a]Monte Carlo derived beta values.

[b]Z test is significant if computed Z value is equal to the absolute value of 3.475.

TABLE 9
FRACTION OF SIGNIFICANT F's USING Z-SCORE ANALYSIS OF VARIANCE AND THERE WAS NO GAIN IN EITHER GROUP

Reliability of Treatment Group Scores	Reliability of No-Treatment Group Scores	Sample Size of Each Group	Size of Gain	Pretest Mean of Treatment Group	Pretest Mean of No-Treatment Group	Fraction of Significant [a] F's	[b] Computed Z-Values
.50	.50	10	0.0	100	100	.038	-1.739
.50	.50	55	0.0	100	100	.058	1.159
.50	.50	100	0.0	100	100	.059	1.304
.70	.70	10	0.0	100	100	.054	.580
.70	.70	55	0.0	100	100	.045	-.725
.70	.70	100	0.0	100	100	.059	1.304
.90	.90	10	0.0	100	100	.049	-.145
.90	.90	55	0.0	100	100	.057	1.014
.90	.90	100	0.0	100	100	.058	1.159
.50	.50	10	0.0	100	80	.057	1.014
.50	.50	55	0.0	100	80	.060	1.449
.50	.50	100	0.0	100	80	.053	.435
.70	.70	10	0.0	100	80	.053	.435
.70	.70	55	0.0	100	80	.039	-1.594
.70	.70	100	0.0	100	80	.050	.000
.90	.90	10	0.0	100	80	.055	.725
.90	.90	55	0.0	100	80	.054	.580
.90	.90	100	0.0	100	80	.055	.725
.50	.50	10	0.0	80	100	.054	.580
.50	.50	55	0.0	80	100	.068	2.607
.50	.50	100	0.0	80	100	.050	.000
.70	.70	10	0.0	80	100	.059	1.304
.70	.70	55	0.0	80	100	.053	.435
.70	.70	100	0.0	80	100	.055	.725
.90	.90	10	0.0	80	100	.001	-7.101
.90	.90	55	0.0	80	100	.003	-6.812
.90	.90	100	0.0	80	100	.005	-6.523

[a] Monte Carlo derived alpha value.

[b] Z test is significant if computed Z value is equal to the absolute value of 3.475.

number of significant F's.

The comparisons of the statistical techniques standard analysis of covariance, analysis of variance on raw gains, standardized analysis of variance, and z-score analysis of variance under nine factor combinations reveals that none of the computed chi-square values was statistically significant. Thus, the four procedures did not produce significantly different results. The statistical techniques analysis of variance on raw gains and z-score analysis of variance under 15 factor combinations indicates none of the computed chi-square values was statistically significant. Thus, the two procedures did not produce significantly different results.

In order to test Null Hypothesis 3, a series of factorial designs were developed for each statistical technique used in this study. The fractions of significant F's were used as the criterion variables in the eight factorial designs where the levels of reliability, sample size, and equality of pretest means were used as factors. The results indicated three statistical techniques produced at least one set of significant interactions (i.e., equality of pretest means and reliability), while factorial experiments using ANOVA on raw gains produced no significant interactions. The results indicated ANOVA on raw gains provided the most realistic results.

Summary

Based on the results of the study, none of the statistical techniques were found to appropriately model reality. In addition, there was no statistically significant difference between the results produced by all four statistical techniques when there was no gain in either group and the pretest means of both groups were equal. Also, analysis of variance on raw gains and z-score analysis of variance produced similar results when there was no gain in either group and

the pretest means were different for both groups. The idea of developing a multiple analyses scheme was not supported for any of the statistical techniques used in this study. Finally, the study did not support the use of the statistical techniques investigated under the quasi-conditions constructed for this study.

Conclusions

The conclusions presented in this section reflect the results found in the study. The standard analysis of covariance, analysis of variance on raw gains, standardized analysis of variance, and z-score analysis of variance were found to be ineffective in providing the appropriate control of Type II errors when there was a gain in either group, regardless of the reliability of the pretest or the size of the sample used. However, all four statistical techniques did provide appropriate control of Type I errors when the pretest means were equal in the no gain situation.

When the frequencies of significant F's for all four statistical techniques were compared where there was no gain in either group and the pretest means were equal, there was no statistically significant difference in the results recorded for each of the four statistical techniques. In addition, the observed frequencies of significant F's for analysis of variance on raw gains and z-score analysis of variance were compared where there was no gain in either group and the pretest means for both groups were different. Subsequently, there was no statistically significant difference in the results recorded for both statistical techniques.

The results obtained from the analysis of each of the four factorial

experiments indicated that equality of pretest means interacted with one or more of the factors (i.e., reliability and sample size) investigated using the fraction of significant F's (alphas and betas), standard analysis of covariance, standardized analysis of variance (no gain situation), and z-score analysis of variance (gain situation). However, when the two factorial experiments involving analysis of variance on raw gain were analyzed, the results indicated no significant interactions among the factors investigated. The combinations of factors used in analysis of raw gains situations produced the most realistic results.

References

Bennett, C. A., & Lumsdaine, A. A. Evaluation and experiment: Some critical issues in assessing social programs. New York: Academic Press, Inc., 1975.

Borg, W. R., & Gall, M. D. Educational research: An introduction. New York: David McKay Company, Inc., 1971.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally and Company, 1963.

Cook, T. D., & Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. Chicago: Rand McNally, 1976.

Cook, T. D., & Reichardt, C. S. Guidelines: Statistical analysis of non-equivalent control group designs: A guide to some current literature. Evaluation, 1976, 3, 136-138.

Coulson, J. E. National evaluation of the Emergency School Aid Act (ESAA): A review of methodological issues. Journal of Educational Statistics, 1978, 3, 1-60.

Cronbach, L. J., Rogosa, D., Price, G., & Floden, R. Analysis of covariance--temptress and deluder? Or angel of salvation? Stanford Evaluation Consortium, 1976.

Gibbons, J. D. Nonparametric methods for quantitative analysis. New York: Holt, Rinehart, & Winston, 1976.

Harris, C. W. Problems in measuring change. Milwaukee, Wisconsin: The University of Wisconsin Press, 1963.

Hendrix, L. J., Carter, M. W., & Hintze, J. L. A comparison of five statistical methods for analyzing pretest-posttest designs. Journal of Experimental Education, 1978, 47, 96-102.

Huck, S. W., & McLean, R. A. Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. Psychological Bulletin, 1975, 4, 511-518.

Kenny, D. A. A quasi-experimental approach to assessing treatment effects in nonequivalent control group design. Psychological Bulletin, 1975, 82, 345-362.

Kleinbaum, D. G., & Kupper, L. L. Applied regression analysis and other multivariable methods. Belmont, California: Wadsworth Publishing Company, Inc., 1978.

Lord, F. M. A paradox in the interpretation of group comparison. Psychological Bulletin, 1967, 68, 304-305.

McKinlay, S. M. The design and analysis of the observational study--A review. Journal of the American Statistical Association, 1975, 70, 503-520.

McLean, J. E. An empirical examination of analysis of covariance with and without Porter's adjustment for a fallible covariate. Doctoral dissertation, University of Florida, 1974. (University Microfilms No. 75, 16,418)

McLean, J. E., Ware, W. B., & McClave, J. T. How analysis of covariance can yield misleading results in educational experiments. Proceedings of the American Statistical Association--Social Statistics Section, 1975, 548-553.

Neel, J. H. A comparative analysis of some measures of change. Doctoral dissertation, University of Florida, 1970. (University Microfilms No. 71-17,198)

Ott, L., Mendenhall, W., & Larson, R. Statistics: A tool for the social sciencies. North Scituate, Massachusetts: Duxbury Press, 1978.

Porter, A. C. The effects of using fallible variables in the analysis of covariance. Doctoral dissertation, University of Wisconsin, 1967. (University Microfilms No. 67-12,147)

Porter, A. C., & Chibucos, T. R. Selecting analysis strategies. In G. D. Borich (Ed.), Evaluating educational programs and products. Englewood Cliffs, New Jersey: Educational Technology Publication, 1974.

Porter, A. C., & Chibucos, T. R. Common problems of design and analysis in evaluative research. Sociological Methods and Research, 1975, 3, 235-257.

Riecken, H. W., & Boruch, R. F. (Eds.). Social experimentation: A method for planning and evaluating social intervention. New York: Academic Press, 1974.

Scheid, F. Theory and problems of numerical analysis. New York: McGraw-Hill, 1968.